The Analysis of Verbal Behavior 1998, 15, 41-56

Protocol Analysis as a Tool for Behavior Analysis

John Austin
Western Michigan University

AND

Peter F. Delaney
Florida State University

The study of thinking is made difficult by the fact that many of the relevant stimuli and responses are not apparent. Although the use of verbal reports has a long history in psychology, it is only recently that Ericsson and Simon's (1993) book on verbal reports explicated the conditions under which such reports may be reliable and valid. We review some studies in behavior analysis and cognitive psychology that have used talk-aloud reporting. We review particular methods for collecting reliable and valid verbal reports using the "talk-aloud" method as well as discuss alternatives to the talk-aloud procedure that are effective under different task conditions, such as the use of reports after completion of very rapid task performances. We specifically caution against the practice of asking subjects to reflect on the causes of their own behavior and the less frequently discussed problems associated with providing inappropriate social stimulation to participants during experimental sessions.

The issue of understanding covert verbal behavior is not a new one: Humans have been interested in the nature and concept of thinking since the days of Aristotle (Sorabji, 1972). As psychology developed into a scientific endeavor, interest in verbal behavior and thinking led to the use of verbal reports as the primary datum in the early 1900s. Some researchers sought to understand human perception of stimuli such as color (e.g., Titchener, 1912) through first training participants and then asking them to verbally respond to such dimensions of perception as *redness*. Needless to say, there are a variety of issues rendering such an approach unacceptable to the scientist, including the possible inaccuracy and invalidity of the reports. For instance, a participant, regardless of the extent or quality of talk-aloud training, could say virtually anything about how he or she perceives the concept of redness. This means that researchers can verify neither the usefulness of the verbalized rules in producing the descriptions of perception nor the descriptions themselves. In essence, the data generated from using this sort of task and procedure are of limited utility in explaining, controlling, and predicting future behavior because they typically constitute only one of a multitude of possible solutions to the problem of perceiving redness. These issues comprise most of what has historically made scientists (behavior analysts and cognitive scientists alike) uncomfortable with the use of verbal reports.

Many argue that at some point, behaviorists rejected the use of verbal reports as data, but this is not true. Watson (1920) recommended the use of verbal reports to study the solution of specific and well-defined problems. Since then, few behavior analysts have used well-defined techniques to collect verbal data. Critchfield and Perone (1990) and Lane and Critchfield (1989) used a preparation that allowed participants to report their perceptions of the accuracy of each response they made in a delayed matching-to-sample task. Wulfert, Dougher, and

We thank K. Anders Ericsson for his helpful comments and criticism of this work.

Correspondence concerning this article should be addressed to John Austin, Western Michigan University, Department of Psychology, Kalamazoo, Michigan 49008 (E-mail: john.austin@wmich.edu).

Greenway (1991) used a talk-aloud procedure to examine the role of verbal behavior in equivalence class formation. Potter, Huber, and Michael (1997) used a similar procedure to examine verbal mediation of selection-based responding in a conditional discrimination task. Wulfert et al. and Potter et al. will each be discussed in more detail later in this paper. Practitioners such as Kent Johnson have reported to utilize a think-aloud problem-solving procedure to help learners at his Morningside Academy (reported in Lindsley, 1996). Although few behavior analysts have reported published studies using protocol analysis, the approach seems to be gaining popularity in the behavioral community (e.g., Delaney, 1997; S. Hayes, 1986; Perone, 1988; Potter et al., 1997; Wulfert et al., 1991).

To conclude, the use of verbal reports as data is not uncommon. Even though both behavior analysts and cognitive psychologists have used verbal reports for years, it is worth reviewing some techniques for ensuring the validity of verbal data. We begin by briefly outlining the position taken by Ericsson and Simon (1993), who have a cognitive psychological approach to the collection and analysis of verbal reports, and then discuss the status of verbal reports in behavior analysis. Next, we describe in detail two behavioral studies that used verbal reports as supplemental data, and explain some methods and considerations involved in obtaining accurate, nonintrusive verbal reports. Finally, we describe some potential pitfalls to avoid in collecting and analyzing verbal reports.

PROTOCOL ANALYSIS IN BEHAVIOR-ANALYTIC RESEARCH

Although psychologists have drastically changed their approaches to studying verbal reports since the early research of Titchener (1912), it is only recently that a comprehensive set of techniques has been compiled to allow the behavioral researcher to reliably study covert verbal behavior. Special techniques for evoking reliable and verifiable verbal reports that do not appear to significantly alter the behavior being studied have been reported by Ericsson and Simon (1993), using the information processing model of thinking. Their proposal, which has been reviewed by behavior analysts (S. Hayes, 1986), is frequently employed by cognitive psychologists (for a short review, see Ericsson & Simon, 1993, p. xi). Although the term we will use to describe their method, *protocol analysis*, is perhaps properly applied only to the actual analysis of verbal data, we will use it also to refer to their general method for collecting accurate verbal reports – or *verbal protocols* – as well as their proposed methods for analyzing those data.

Protocol Analysis: Definition and Standard Method

Protocol analysis is a set of methods for obtaining reliable information about what people are thinking while they work on a task. Although the term applies to a variety of methods for obtaining and analyzing verbal reports, we focus primarily on the method preferred by Ericsson and Simon (1993), namely the use of concurrent verbal reports. With this method, rather than asking the participant questions (as might be done in an interview situation), the experimenter simply asks the participant to "think aloud" and verbalize his or her thoughts as if the participant were alone talking to him- or herself. The procedure is not conversational, and attempts to minimize social verbal behavior. Although the result sometimes includes idiosyncratic verbalizations and may not always be easily interpreted, the procedure reduces biases of other types. For reference, we have reproduced a standard set of participant instructions from Ericsson and Simon (1993) in the Appendix. After seeing the basic instructions, the participant is given practice at talking aloud while solving some simple problems, such as mental arithmetic problems, during which time the experimenter prompts the participant to talk out loud as if he or she were in the room alone. Once the experiment itself begins, the experimenter's role is only to prompt the participant to "please keep talking" if he or she becomes silent.

At the end of the experiment, the participant's verbalizations are transcribed into text from the tape recording. Unfortunately, there is no hard-and-fast set of rules for

doing protocol analysis; different experimenters use slightly different variations on the same basic procedure. Analysis typically proceeds by segmenting the protocols into blocks of text. The size of these blocks depends on the level at which the data become orderly; the experimenter must decide ahead of time what the variables of interest are, and these are often based on the theory or phenomenon being studied (e.g., Wulfert et al., 1991, as discussed later in this paper, coded according to the theory that relational responding facilitates equivalence formation). For most behavior analysts, we assume that relatively specific descriptions of individual behaviors or descriptive categories such as Skinner's (1957) basic verbal operants (as used by Potter et al., 1997) will be coded in efforts to identify functional relationships.

Once the transcribed protocols have been segmented into blocks, the blocks are then randomly reordered and presented out of context to coders who rate each block based on a scheme devised by the experimenter. Segmenting is a procedure used to create individual statements from long and continuous transcripts. Behavior analysts have sometimes dealt with this through task design rather than spending time deciding how to segment transcripts. For instance, both Potter et al. (1997) and Wulfert et al. (1991) divided protocols using each experimental trial as the start and end of a segment. After segmenting, coders should receive the segments with only the necessary context to make categorical decisions about the data. The coded blocks are then reassembled in their original order for analysis, followed by some check on interrater reliability.

Status of Verbal Reports in Cognitive Psychology

The methods of Ericsson and Simon (1993) are based on the information processing view, which we review here briefly in order to clarify the theoretical status of protocols in behavior analysis. The information processing model, as used by cognitive scientists, assumes that there is a set of underlying, hidden cognitive mechanisms that generates all potentially observable human behavior. These cognitive processes can be measured indirectly through standard experimental means, such as by recording reaction times or by counting correct and incorrect responses. This theoretical perspective implies that any observed verbalizations are "produced by the same cognitive processes that produce more traditional performance data" (Ericsson & Simon, 1993, p. xii). Knowing this, one could theoretically predict under what conditions verbal data would be accurate or inaccurate given a particular cognitive theory. The special techniques we will describe for protocol analysis were originally designed to conform to such a theory.

The specific theory that Ericsson and Simon (1993) adopt is based on several assumptions. First, cognitive processes are viewed as *sequences of internal states* that are somehow transformed by information processes. These states constitute the things that a person is aware of at a particular time and are represented as symbols, such as words, ideas, or images. The notion that these states are represented is important to the theory, because that is used to explain why they can later be reported – at some level, the person has encoded the world in terms of these symbols. The particular processes used to move from state to state might themselves be impossible to report directly, and in fact the person might not even be aware of them at all.

The next assumption is that there are *several memory systems,* each having its own processing and capacity characteristics in which information (in the form of symbols) is stored, retrieved, and processed (Ericsson & Simon, 1987; Simon, 1979). According to the theory, there are at least two memory systems, a short-term system used for storing the states temporarily and a long-term system used for storing the symbols for later use. Symbols in short-term memory are virtually immediately accessible, whereas those in long-term memory can only be retrieved using an information process that retrieves them using a "retrieval cue" – another symbol in short-term memory.

Taken together, these assumptions imply that the most reliable verbal reports will be those taken concurrently, that is, during

performance of the task. Retrospective reports may also be valuable, but they are subject to errors in retrieving information about what happened. An organizing principle for Ericsson and Simon's (1993) approach is that only the contents of the short-term memory system are accessible through the talk-aloud procedure. This means that whereas participants can reflect on prior (non-short-term memory information) events with limited reliability, they can reliably report only information that has not yet left the short-term memory system. Verbalizations generated under the conditions specified by Ericsson and Simon (1993) are directly retrieved from the information heeded as tasks are completed. This information, they argue, consists of stimuli that are in short-term memory during task completion (even when not verbalized) and therefore do not introduce additional inferences or thought processes on the part of the participants.

Ericsson and Simon (1993) argue that the constraints of the information processing model must be understood and accepted in order to verify that protocols will be accurate. But how important is it for behavior analysts to accept these assumptions when using protocol analysis? How can one understand protocol analysis from a behavior-analytic perspective?

Status of Verbal Reports in Behavior Analysis

Although many stimuli and responses are observable, much of both the behavior involved in thinking and its context are inapparent to observers at the time it occurs (L. Hayes, 1994). Even though all overt behavior is, in theory, traceable back to some environmental event (Shimoff, 1984), it also seems clear that covert verbal behavior is often an intermediate link in this causal chain (Potter et al., 1997). Because behavior analysts have traditionally studied only those responses and stimuli that are apparent, the nature of thinking and covert verbal behavior presents a special problem. Skinner (1974) suggested that observing covert verbal behavior is not difficult, but he did not explain how to do it reliably. From a behavioral perspective, it is logical that in some cases the use of concurrent talk-aloud protocols could provide a procedure for making apparent some of these inapparent stimuli and responses, allowing appropriate analysis of establishing operations, antecedents, and consequences that control verbal behavior and therefore effective manipulation of the behavior and its context.

More specifically, our assertion is that protocol analysis, as a set of techniques, can be used to bring critical covert verbal behavior to the overt level. If we can accept the assumption that, under certain circumstances, it is possible to confirm such a covert-to-overt transition, then what follows is easily accepted.

Behavior analysts have long argued against the use of hypothetical constructs such as memory storage systems, and have also argued that the memory storage metaphor is not necessary for verbal reports to be accurate and for protocol analysis to be effective in identifying the self-statements participants use (S. Hayes, 1986; Wulfert et al., 1991). Hayes described an example in which a participant was asked to say aloud what he or she was already saying privately while engaging in a task. Under these conditions, Ericsson and Simon (1993) would argue that verbalization does not affect task performance. Hayes further suggested that there are at least two possibilities of interpretation for the resulting verbal report. The verbal report could be generated by a response system independent of the one producing the task performance and therefore represents "an additional irrelevant response" (Hayes, 1986, p. 357). Alternatively, the task performance could be unaffected by verbal reporting because the equivalent events "have already influenced task performance" (Hayes, 1986, p. 357). Perhaps the behavior is already under the control of antecedent verbal stimuli, and the verbal report is a tact of this antecedent stimulation.

If the verbal report indicates an additional irrelevant response, it is of little interest to behavior analysts, because behavior analysts most frequently focus on the manipulable variables, either environmental or behavioral, that are associated with behavior. However, as S. Hayes (1986) argued, why does task-relevant verbalization disrupt task performance in some cases but not in other situations?

If the verbal report represents a tact of antecedent verbal stimulation, it may be very useful to behavior analysts. Wulfert et al. (1991) used the results of protocol analysis to supplement their analysis by demonstrating a functional relationship between participants' verbalizations and the target responses. The experimenters employed a dual-phase method. During the first phase, think-aloud protocols were collected, and responses correlated with successful performance on the task and with failed performance on the task were identified. Two different groups of participants were then trained using the strategies of the successful and unsuccessful participants from the first phase. Participants trained in the successful strategy were more successful at the task than those trained in the unsuccessful strategy.

Clearly, behavior analysts do not need to accept Ericsson and Simon's (1993) information processing assumptions in order to successfully identify functional relationships between antecedent verbal stimuli and task performance. Indeed, we agree with S. Hayes (1986) and therefore take the position that it is not necessary to achieve and demonstrate a one-to-one correspondence between verbal reports and the contents of short-term memory, as Ericsson and Simon argue (and it is therefore also not essential for us to argue for or against the existence of short-term memory). Rather, as behavior analysts, we are more interested in demonstrating and establishing the functional properties of specific statements, rules, and strategies in the completion of certain tasks.

TWO EXAMPLES OF BEHAVIOR-ANALYTIC STUDIES USING TALK-ALOUD PROCEDURES

A few behavior analysts have used talk-aloud procedures to help participants produce supplementary data that are useful in suggesting answers to experimental problems. Wulfert et al. (1991) investigated a potential source of individual differences in equivalence class formation, and Potter et al. (1997) investigated the mediational role of verbal stimuli in learning conditional discriminations.

Untrained conditional relations often, but not always, emerge when humans are taught conditional discriminations to stimuli within a related set (Sidman & Tailby, 1982). When learned, these untrained and trained relations are said to participate in a *stimulus equivalence class* (Sidman, 1971). Some stimulus equivalence experiments have shown that equivalence is difficult to train in some participants (e.g., Devany, Hayes, & Nelson, 1986). The reasons for these individual differences in equivalence class formation are "not well understood" (Wulfert et al., 1991, p. 489). Logically speaking, individual differences of this type are difficult to explain based only on behavioral outcome measures (e.g., matching-to-sample accuracy) because these measures are not always indicative of the specific stimuli controlling behavior. That is, when a participant correctly (or incorrectly) selects a stimulus as matching another stimulus, experimenters cannot observe all stimuli to which the participant is responding. Some have argued that selection-based responding (e.g., matching to sample) is sometimes mediated by verbal responses (Michael, 1993). Collecting and examining verbal reports could reduce the amount of time spent in additional careful experimental efforts to uncover the verbal stimuli controlling the matching-to-sample response.

Wulfert et al. (1991) used participant verbalizations as supplementary data to suggest that individual differences in equivalence class formation are a result of particular self-statements. Participants were trained and tested during a 2-hr session. They were instructed to "think aloud during the entire experiment" (Wulfert et al., p. 491). The experimental session began by training participants to talk aloud on an arithmetic problem. During training, when posed the question, "How much is 127 plus 35?" if participants simply stated the answer without talking aloud, the experimenter modeled appropriate responding by saying, "Assume the problem is 123 plus 66. To solve it, I will think 123 plus 6 makes 129, plus 60 makes 189. Now here's another problem. Solve it thinking aloud" (Wulfert et al., p. 491). This continued until participants modeled the experimenter's talking aloud on two consecutive trials.

When the experiment began, participants engaged in a computerized matching-to-sample task with eight experimental stimuli. For the first 3 to 5 min of this phase, the experimenter stayed in the room and prompted, "Don't forget to think out loud" (Wulfert et al., 1991, p. 491) when participants did not verbalize during any two consecutive trials. The A-B, A-C, and A-D discriminations were trained successively to criterion. The computer program generated a tone at the start of each trial and a double tone at the start of every 10th trial, so that tape-recorded verbalizations could be linked to specific trials. The verbalizations that occurred during each trial were transcribed and coded into four categories. (a) *Relational responding* included statements referring to the relationship between two stimuli; (b) *common physical features* included statements relating pairs of stimuli by their nonarbitrary aspects; (c) *stimulus compounds* included statements indicating visual integration of sample and comparison stimuli; and (d) *other* included statements not coded in the existing categories.

Wulfert et al. (1991) found that individual differences in learning conditional discriminations were correlated with different vocalizations from the classification system described above. More specifically, of the 10 participants, the 3 who did not demonstrate equivalence spoke more about the common physical features and the stimulus compounds, whereas those who achieved equivalence spoke more about the relationships between sample and comparison stimuli. These data suggest a correlation between verbally responding to relationships between stimuli and equivalence class formation.

To test this hypothesis, the experimenters conducted a second study in which they trained relational responding in one group and compounding in another. The same computerized task was used for this study, but the stimuli and participants were changed. Participants performed equally well during training, but during symmetry and equivalence tests, 6 of 7 participants in the compounding group failed, whereas 6 of 7 participants in the relational responding group performed to criterion.

Another study conducted by Potter et al. (1997) investigated the possibility that, for individuals with strong verbal skills, verbal behavior plays a mediational role in performance on selection-based tasks. Two types of verbal behavior are selection-based and topography-based verbal behavior (Michael, 1985). Verbal behavior is said to be selection based when an individual identifies a verbal stimulus through pointing, touching, or otherwise physically gesturing. Experimental preparations teaching conditional discriminations to humans are usually selection based because participants are required to point and click using a mouse, to select a card, or to touch an object on a computer screen. Alternatively, topography-based verbal behavior involves speech or otherwise unique topography to provide feedback to the speaker and to which the listener can respond. Researchers have shown that topography-based methods of instruction are more effective than those using selection-based responding, resulting in faster acquisition and higher accuracy rates, especially for those with limited verbal skills (Hodges & Schwethelm, 1984; Sundberg & Sundberg, 1990). In studies with highly verbally skilled individuals, the fact that performance differences are small between selection- and topography-based verbal behavior (Bristow & Fristoe, 1984) has led some researchers to hypothesize that selection-based verbal behavior is often mediated by topography-based verbal behavior (e.g., Lowenkron, 1991).

Potter et al. (1997) instructed 4 college students to talk aloud while they were engaged in a conditional discrimination task. Participants were required to choose among 12 stimuli to find the correct comparison stimulus to a sample in the center of a computer screen. Verbalizations were tape recorded, transcribed, and coded using two of Skinner's (1957) basic operants, the tact and the intraverbal, as categories. Rather than looking for differences between participants in the end result of equivalence class formation, the experimenters coded and compared verbal responses immediately preceding correct and incorrect selections during training versus those made during testing. They found that a high percentage of correct selections during testing were preceded by

the same tacts to the sample stimulus that were used during training conditions and the same intraverbal that was emitted in response to the choice (e.g., comparison) stimulus during training. In addition, the most frequent type of statement that preceded incorrect responses during testing was the same tact to the sample stimulus as was used during training and an intraverbal (or no intraverbal) different from that used during training in response to the choice stimulus.

The results indicated that while engaging in the conditional discrimination task, when participants verbalized those tacts and intraverbals used during the training phase, they performed better than when they did not. This provides a within-subject analysis and extension, using a different coding scheme, of the Wulfert et al. (1991) study described above. Both studies, however, effectively utilized protocol analysis techniques to collect and make sense of verbal data. Both studies used the verbal data as a supplement to their primary datum, and both used their primary datum to demonstrate the functional significance of the specific types of verbal statements.

We will now discuss and summarize methods, some of which were used by both of the above-described studies, for obtaining accurate and nonintrusive verbal report data.

METHODS FOR OBTAINING ACCURATE[1] AND NONINTRUSIVE VERBAL REPORTS

Avoiding Reactivity and Obtaining Reliability of Coding

In the case of verbal reports, the following question often arises: "Did requiring the participant to think aloud during the task in some way alter his or her performance on the task?" In their book, Ericsson and Simon (1993) spend a great deal of time arguing that protocol analysis does not cause reactivity in most nonautomatic (i.e., contingency shaped; see S. Hayes, 1986, for a behavioral view of the cognitive term *automatic*) tasks that require short-term memory processing. Chapter 2 of Ericsson and Simon's book is devoted to this issue, and provides a good summary of early tests of reactivity in talk-aloud studies. Others have attempted to demonstrate that concurrent verbal protocols do not cause reactivity for tasks in specific domains (such as management; Schweiger, 1983).

There is clear evidence that certain kinds of verbal reports sometimes do in fact produce changes in task performance. Perhaps the most dramatic of these are what Ericsson and Simon (1993) called Type 3 verbalizations, in which participants are asked to introspect on their performance or provide reasons for their behavior. For example, studies have found that forcing participants who have not previously solved the Tower of Hanoi task to give reasons for their moves improves their performance on the task relative to controls (e.g., Ahlum-Heath & DiVesta, 1986; Gagne & Smith, 1962; Wilder & Harvey, 1971).

Various behavioral researchers have studied reactivity to self-reporting in applied (see Nelson, 1977, for a review of self-monitoring in clinical settings) and experimental (Critchfield & Perone, 1990) settings and have found self-reporting to affect performance. For example, Critchfield and Perone found that when participants were asked to report the perceived correctness of their selection in a computer-aided task, self-reports were less accurate when the target response was under greater time pressure, and that performance was slowed during self-report conditions.

When the standard think-aloud instructions are used, however, Ericsson and Simon (1993) found that concurrent verbalization seems not to significantly alter participants' behavior. In particular, they claim that when the specific techniques of protocol analysis are followed, as long as appropriate instructions are employed, the only possibly substantial effect of verbalization on problem solution is that when participants think

[1]As the preceding discussion implies, by accurate we do not mean that the verbal reports should or must correspond to their internal states or that such states exist. Rather we simply mean that accurate verbal reports are accurate in the sense that they are functionally equivalent to those rules used by other participants who correctly complete the task but do not verbalize overtly.

aloud they take longer to complete the problems. In support of their claim, they review numerous studies that directly compare performance of think-aloud participants with silent controls. Across different tasks, studies of thinking aloud have found no differences in behavior between think-aloud participants and silent controls. For example, Newell and Simon (1972) took detailed verbal protocols from 7 individuals trying to generate propositional logic proofs. For two problems, they compared the specific observed proof steps with those discovered by a group of participants in a silent condition (conducted earlier at a different site), and found essentially no differences between the proofs generated by participants in the talk-aloud condition and those generated in the silent condition.

Even when participants report the subjective experience that protocols are altering their performance, they may not be. Karpf (1972), for example, compared the performance of participants talking aloud with those who performed silently on a discrimination learning task. Participants in each group were matched in a pairwise fashion based on performance on 10 pretest problems. Following 15 experimental trials, the experimenter presented five trials in which all participants worked silently and five in which they all thought aloud, to detect any lingering effects of thinking aloud. Although there were no differences in terms of solution accuracy, whenever participants were asked to think aloud they took significantly longer to complete the problems. After the experiment, participants were asked whether they thought thinking aloud helped, hindered, or had no effect on their performance. Although those who indicated that thinking aloud hindered them performed significantly worse than those who indicated that thinking aloud had no effect, the fact that the former group's poor performance did not improve during the five final silent problems calls into question the validity of their claim. It seems more likely that participants took the opportunity to blame the situation rather than their own skills.

Other studies have demonstrated that when protocol analysis guidelines are followed, talking aloud does not interfere with task completion. For example, Bower and King (1967) tested for and found effects of verbalization only as an increasing function of the number of irrelevant dimensions of the problem. Karpf and Levine (1971) found no effects of verbalization on performance in a discrimination learning task. Brehmer (1974) found no effects of having participants verbalize the rule they used to predict answers in a cue-probability task. Some of these studies did not even use techniques as conservative as recommended in this paper, by Ericsson and Simon (1993), or by S. Hayes, White, and Bissett (1998).

Even tasks that require perceptual-motor or visual processes remain unaltered by think-aloud instructions. However, note that in such cases the protocols may be difficult to interpret; frequently, participants say little beyond describing the physical moves they are making. From a behavioral perspective this makes sense because often verbalization is neither required nor functional when engaging in perceptual-motor or visual tasks. It becomes functional to do so only after the experimenter instructs a participant accordingly.

Despite the similar conclusions of the more than 40 studies reviewed by Ericsson and Simon (1993), we nevertheless recommend checking for reactivity when introducing a new experimental task, because the issue of reactivity is an empirical, rather than a logical, one (Russo, Johnson, & Stephens, 1989). In addition, the functional effects of verbalizations can be confirmed through further experimentation (as in Wulfert et al., 1991).

S. Hayes (1986) recommends that researchers rule out reactivity by including a silent control group (no talking aloud) to compare their terminal performance to that of the talk-aloud group, a technique similar to those used by cognitive psychologists. We recommend that if it is repeatedly demonstrated that there is no reactivity for a given specific task (i.e., no differences are found between talk-aloud and silent control terminal performance), then one could assume that these conditions would hold for future experimentation. This does not hold for entire domains, however, because tasks may vary widely within a given domain. For ex-

ample, in the domain of mathematics, we cannot say that because there is no reactivity for participants talking aloud while solving addition problems that this will hold for those solving advanced calculus problems. The task, its dimensions, and the instructions provided must be carefully taken into account.

Another threat to the internal validity of protocols is unreliability of coding. This is different than the unreliability of *reporting* in that coding occurs, as a way of summarizing the data in quantitative terms, after reports are collected and transcribed. Many researchers have published papers in which they used verbal reports and had one person code all segments. This is unacceptable because when human judgment (i.e., in coding protocols) is involved, human error is possible. Ericsson and Simon (1993) recommend coding about 10% of the total observations for reliability, but we take a position closer to the *Journal of Applied Behavior Analysis* standard of 25% of observations.

Wulfert et al. (1991) had 9% of the transcripts from the first experiment checked for accuracy from the tape recordings and checked reliability of coding for 7% of statements across 4 of 14 participants during the second experiment. Potter et al. (1997) checked reliability of coding for 100% of statements, reporting the scores for the group, for individuals, and for one of the seven coding categories. When reliability coders do not check 100% of statements, reliability coding should ideally be randomized across groups and stratified by coding category (there are computer programs for Macintosh® and PC computers that can assist with this randomization; see, e.g., Crutcher, Ericsson, & Wichura, 1994).

Instruction and Practice

Ericsson and Simon (1987) reviewed some instructions used by other researchers (e.g., Duncker, 1926) and conclude that the instruction should be brief and should tell the participant to "think aloud"[2] while making reference to a process familiar to the participant. A reasonable example of this would be, "In this experiment we are interested in what you think about when you solve performance problems. We would like you to think aloud as you solve the problem, simply speaking aloud as if you were alone and thinking to yourself."

Ericsson and Simon (1993) also recommend that experimenters instruct the participant not to try to explain anything to anyone, but to rather simply think out loud, not justifying anything that they say. Behavior analysts have shown that participants often have little success at describing the contingencies that control their behavior (Catania, Shimoff, & Matthews, 1989), and this is no different; if we ask participants to explain why they respond the way they do, we are asking them what is causing their behavior. This is important because it has been demonstrated that participants will create explanations for their behavior and explain why they said what they said (we will return to this point later when we discuss frequent problems with verbal reports studies). Because this is not the focus of our search when conducting protocol analysis, we wish to minimize such behavior.

After the initial instruction, it is essential that experimenters provide practice in thinking aloud. Ideally, this would be achieved by using a task that is very similar to the experimental task. When the experimental task is difficult for the participant, the experimenter may choose to include several practice tasks, each requiring successively more effort on the part of the participant. For example, Austin (1996) used as the experimental task three extensive business management cases. Because the extent to which participants had previously engaged in thinking aloud was unknown, participants were given practice using (a) a simple multiplication problem, (b) an exercise requiring a detailed description of the windows in the participants' parents' house, and then (c) a case describing a military performance problem (taken from Boreham, 1986).

The practice session is a time in which the experimenter should shape participant responding while the task solution is ongoing. For instance, when given a multiplication

[2]In this paper, we refer to the procedure as *talk aloud*, partly because *thinking* is a word loaded with connotation that we do not intend. Alternatively, in reference to participant instructions, we always use the term *think aloud* because the phrase effectively prompts participants to appropriately talk aloud during the task.

problem such as "What is 24 times 6?" some participants will pause and then give the final answer. The experimenter should at this point explain the think-aloud procedure again: "Instead of giving only the final answer, I want you to report each step of the problem solution." Wulfert et al. (1991) used a similar procedure whereby the experimenter modeled appropriate verbalization when participants simply responded with the final solution rather than talking aloud while solving the problem.

Tasks Appropriate for Study Using Concurrent Reports and Alternative Methods for Making Inapparent Behavior Apparent

Not all tasks are appropriate for evoking accurate and valid concurrent verbal reports. In particular, some tasks occur over time intervals that are too short to allow effective concurrent verbalization, as in memory research (Chase & Ericsson, 1981, 1982; Ericsson & Polson, 1988; see also Delaney & Austin, 1998). Critchfield and Perone (1990) prompted self-reporting of perceived accuracy immediately after participants made a choice in a delayed matching-to-sample task. In cases like those above, retrospective reports may be useful. Although retrospective reporting is known to often cause fabrication of information and verbalization of inaccurate information in long tasks (e.g., Reber & Lewis, 1979), there is no such evidence for tasks that last between 0.5 s and 1 s when instructions to remember the specific thoughts without commenting on them are used (Ericsson & Simon, 1993).

Retrospective reporting techniques may also be useful when studying experts, when in many cases a great deal can be accurately reported about performance on a task in their domain of expertise (e.g., Chase & Ericsson, 1981, 1982). In practice, experts must make responses long after the relevant stimuli are no longer available (Ericsson & Delaney, in press). In cases such as chess games, key behavior cannot be evaluated by the player until the game is actually won or lost. If the relevant responses could not be recalled accurately (assuming they were not transcribed as in tournament play), it would be impossible (or prohibitive) to improve at such tasks.

Among tasks in which concurrent reports can be collected, Newell and Simon (1972) characterized problems as existing on a continuum of well defined to ill defined. Well-defined problems are those in which a single correct answer is verifiable either empirically or through using accepted rules and procedures to determine the answer. Mathematics and physics are domains in which there are many well-defined problems. Simple arithmetic problems are the most obvious examples: There is only one correct answer to 4+4. In contrast, ill-defined problems are those having more than one possible correct answer. Problems in the social and behavioral sciences are often thought to characterize ill definition. The example of Titchener's (1912) participants' thinking aloud about their perceptual experience of redness is a case in point. Chess is an example that, because of its complexity, is not as well defined as mathematics problems but is better defined than perceptual experience. We are not arguing that protocol analysis can only be used in studying well-defined tasks, but rather that one needs to consider the definition of the task and arrange the experiment accordingly. If the definition of the task is not considered in planning the experiment, one may find that there is no reliable method for comparing participants' answers to the correct standard (as in the perception example above).

Expert chess strategy has been widely studied using protocol analysis techniques, so we will use a brief example from that domain. One approach experimenters have used in better defining some components of expert chess play is to essentially create a situation in which there is one best answer. Called the "best next move," this approach presents a player with part of a board or a series of partial boards in midgame and asks the player to think out loud about what the best next move would be (de Groot, 1965). The correct answer can be verified through having a computer play out all possible sequences given the selected move. An approach like this could be used for other tasks to identify contexts that make responding more or less difficult for participants. For example, Wulfert et al. (1991) suggested that

equivalence is facilitated when symmetry probes are presented before equivalence probes. Experimenters could vary the order of presentation of symmetry versus equivalence probes in testing equivalence while participants talk aloud to better understand why this occurs.

The "best next move" is effective for chess, but we are often interested in studying applied problems with less definition. The key is always to have some verifiably correct answer, even if you must create an artificial task that resembles your actual task of interest. This is one way to deal with the problem of verifiability: When there is a correct answer, it is clear if the participant answers correctly or not. (Creating a task analysis is another method of ensuring verifiability, and it is described below.)

For example, Austin (1996) studied the ill-defined task of expert, managerial, and novice solutions to organizational performance problems. Because organizational performance problems often seem to have several potential solutions, which solution is "correct" is often an empirical question. Austin actually solved an applied problem, documenting along the way all of the information needed to solve the problem, summarized this information, and asked participants to solve it when given only the initial presenting problem. This allowed participants to ask questions about the case and to talk aloud about the answers they received from the experimenter.

The question of correctness is solved in this instance, but not entirely. When applied problems are solved, there is one solution that was effective in reality, but there may well be other solutions that may have been effective to varying degrees. This procedure allowed the experimenter to evaluate participant responses based on an actual solution of the problem, and to argue against the inclusion of certain other answers on the basis of empirical work of other authors. For example, to solve a problem of absenteeism, applying appropriate antecedents and consequences may be effective, but sending employees to training has a low probability of effectiveness. Modifying the task in this manner allows researchers to study how experts solve applied problems so that novices may be taught the skills they need to be effective (in a manner similar to the Wulfert et al., 1991, research).

The above considerations in task design deal with evaluating terminal performance, or the final "answer" participants give. An a priori task analysis is recommended by Ericsson and Simon (1987, 1993) as another method of discriminating between "correct" and "incorrect" protocols. The clearest example of the utility and construction of an effective task analysis is in arithmetic. For example, if we ask participants to think aloud while answering the question, "What is 125 plus 25?" the participant has several possible sequences through which to obtain the correct answer:

(a) 100 + 25 + 25 = 150

(b) 25 + 25 = 50 + 100 = 150

(c) 100 + (25 + 25) = 100 + 50 = 150

(d) 5 + 5 = 10; 2 + 2 + 1 (carried from the 10) = 5; 1 + 5 + 0 = 150

There are more possibilities; however, there are only a finite number of correct sequences (i.e., solution paths) useful in determining the correct answer, and herein lies the utility of the task analysis. Ericsson and Simon (1993) convincingly argue that if we give this task to a participant and the participant's protocol indicates that he or she used one of the processes hypothesized in the a priori task analysis, it is more parsimonious to assume that this was the path the participant used. If the participant did not know the other possibilities in the task analysis, it is illogical to argue that the participant verbalized one correct thing but thought another correct thing. In addition, the fact that the participant's protocol agrees with the task analysis indicates that he or she did not simply invent the protocol, because it is again less parsimonious to assume a participant to be intentionally deceiving experimenters through giving the correct verbalization while engaging in a different process to get the correct answer.

In practice, tasks are rarely as well defined as the arithmetic problem above. However, when using more ill-defined tasks, the task analysis provides a method of defining the "correct" process. Task analysis can be

thought of as the process equivalent to effective task design; the experimenter needs to define a priori every possible way that participants can achieve the correct answer. As problems become more ill defined, however, the problems with this approach become clear: For the most ill-defined problems, there may be literally millions of possible solution sequences. One way to avoid the overwhelming task of identifying all possible processes is to specify an *optimal* solution path task analysis and measure the degree to which participants depart from it. Although this is far from a perfect solution, it is more rigorous than not using a task analysis at all.

Another approach may be to vary the required verbal response so that it is extremely simple. For instance, Critchfield and Perone (1990) eliminated the need for talking aloud and the related problems in their study of verbal self-reporting. In a delayed matching-to-sample task, they used a computer program that occasionally prompted participants immediately after they selected a stimulus and asked them to report if they thought they made the correct or the incorrect selection. Although verbalizations were not collected, the Critchfield and Perone study isolated two responses that were each verifiable and therefore did not need a formal task analysis in order to code the responses.

In a study described earlier in this paper, Wulfert et al. (1991) collected verbalizations but defined the coding categories a priori according to their theory of relational responding. Wulfert and colleagues were not interested in understanding a sequential process of verbal stimuli per se. Rather, they found the supplemental evidence they required in the *content* of participant vocalizations, irrespective of the order in which they were spoken.

MANIPULATIONS THAT INCREASE THE PROBABILITY OF REACTIVITY IN VERBAL REPORTS METHODOLOGY

Obviously, participants' verbal behavior can be modified through interaction during the experimental sessions. Therefore, we generally recommend that interaction be kept to a minimum, with the experimenter standing behind the participant if at all possible (as in Potter et al., 1997). Of course, the experimenter must be able to respond to questions from the participant in order to ensure task-relevant behavior (as in Wulfert et al., 1991). The standard instructions are geared towards selectively reinforcing verbalizations without triggering other social responses such as explanations; prompts to "please think aloud" after silences and similar behavior are of course not frequently encountered in everyday conversation.[3]

Some other kinds of prompting can serve as discriminative stimuli for task-irrelevant verbal behavior, or worse, can actually modify the task-relevant behavior in unpredictable ways. Anyone who has worked extensively with protocols can attest that they appear to contain a great deal of data about process, but it is sometimes difficult to determine what a particular fragmentary statement means. As such, it is often tempting to ask participants to clarify their verbalizations or to comment on their thinking process, either during the session or in a later session. For example, recent work on acquisition strategies in paired associates learning tasks were discussed during a panel discussion (Potter, 1997) in which researchers reported asking participants to "explain the meaning" of some of their verbalizations at the end of sessions. It is generally believed that such instructions can produce misleading responses on the part of the participants (Ericsson & Simon, 1993), a fact that is well known to behavioral researchers who have heard participants verbalize "rules" that do not match their contingency-shaped behavior (see Skinner, 1974, for a lucid argument of this claim). The talk-aloud procedure does not ask participants to produce any additional behavior; it simply asks them to overtly report what they are already doing covertly. There are several known studies supporting the claim that asking participants to verbalize rules can interfere with

[3]This also suggests that the experimenter's friends and colleagues are less than perfect participants in verbal reports studies, because the experimenter himself or herself can serve as a powerful discriminative stimulus for non-task-relevant verbalizations.

contingency-shaped behavior or indicate insensitivity to contingencies (e.g., Hendrix, 1947; Rommetveit, 1960; for reviews and discussion, see Catania et al., 1989, and Ericsson & Simon, 1993, pp. 102-104).

Similar problems occur when asking participants to give reasons for their behavior. Often, participants do not have any special insight as to the causes of their own behavior, and their inferences about the reasons why they did particular things may even be different depending on what stimuli are available to them. For example, Storms (1973) videotaped several participants conversing and then asked them their reasons for making particular comments. He then showed them the videotape and asked them the same questions. His finding was that participants attributed their behavior more to situational factors in the former case (when conversing) but to internal variables in the latter case (when watching themselves converse).

Potter et al. (1997) reported that in a post-experimental session, they asked participants about the strategies they had used. The experimenters reported that some participants could not remember the nonsense syllables they had used during the sessions, although participants seemed to be convinced that verbal behavior mediated their selections. Other participants seemed to edit their explanations, "as they said they were goofy or something of that nature" (Potter et al., 1997, p. 53). Data such as these suggest that the participants are inferring the causes of their behavior from the currently available stimuli (for a more detailed, although non-behavior-analytic, discussion of this problem, see Nisbett & Wilson, 1977).

In more extreme cases, it is possible to influence the verbal behavior of participants in ways that could lead to erroneous conclusions about the actual topology of behavior in different tasks. Research in social psychology suggests that leading questions can produce erroneous "memories" on the part of participants, so that at a later time verbal prompts to remember a fictitious event will cause them to recall, as fact, an event that never actually happened. In their now-classic study, Loftus and Palmer (1974) showed a film of a traffic accident to participants and then asked them a series of questions. The critical manipulation was that one of these questions was either "How fast were the cars going when they *smashed* into each other?" or a slightly reworded version that used a different verb phrase (e.g., *hit, collided with,* etc.). Not only did the different questions produce different answers, but they each produced lasting changes in behavior. For example, participants who were asked about cars that smashed into one another were significantly more likely than those asked about cars that collided with one another to recall broken glass a week later, when in fact there was no broken glass in the film. An impressive collection of related studies has been conducted by Loftus and colleagues (for a short review see Loftus & Hoffman, 1989). Many of these studies show that even very elaborate "false memories" can be created with the proper stimulus control and reinforcement. Some participants reportedly refuse to acknowledge that the "memory" was produced by the experimenters, even when told about the procedure during debriefing (Ceci & Bruck, 1993).

There is good evidence that the confabulations produced by participants in these studies introduce permanent changes in participants' remembering behavior, and that it is possible not only to distort memory but also to "implant" entire systems of memory (Loftus, 1996; Manning & Loftus, 1996). As an example, Loftus and Pickrell (1995) provided 24 participants with descriptions of three true events and one false event each had supposedly witnessed with a close family member when the participants were children. They found that even after 2 weeks, some participants refused to believe that the "implanted" memories were indeed false.

In conclusion, inappropriate interaction with participants can produce erroneous conclusions about the topology and causes of behavior as well as unintended changes in the behavior outside of the laboratory. In particular, participants should not be given potentially leading information that could bias their subsequent responses or be asked to provide reasons for their behavior.

CONCLUSION

We reviewed some methods for evoking reliable and valid verbal reports based on those suggested by Ericsson and Simon (1993) and S. Hayes (1986). Although psychologists since Watson (1920) have suggested the use of verbal reports, only recently have these techniques been explicated. The assumptions of the information processing model outlined by Ericsson and Simon (1993) have been described as requirements for obtaining reliable and valid verbal reports. However, it is not necessary to accept the constraints of the information processing model. Rather, one need only accept that it is possible, arguing from a behavior-analytic perspective, to obtain such reports based on established laws of behavior. In addition, it is possible to use protocol analysis techniques to conduct a more thorough experimental analysis of behavior. This can be accomplished through identifying the establishing operations, antecedents, and consequences that control verbal behavioral processes and demonstrating their functional relationship to the target behavior.

Based on these laws of behavior, several issues arise in designing experiments and tasks as well as in providing appropriate instruction to participants. Reliable and valid reports may be obtained if participants are provided with practice exercises; if they are instructed to think aloud but not to explain their behavior; if the task does not induce reactivity; if tasks are designed to have a single best solution; and if a priori task analyses are constructed for the purpose of evaluating protocols. In addition, experimenter prompting should be kept to a minimum and reports should be concurrent whenever possible, because it has been demonstrated that recall is sometimes changed in a lasting way when such conditions are not met. If each of these conditions are met, verbal reports can serve as a useful tool in the study of covert verbal behavior by behavior analysts.

REFERENCES

Ahlum-Heath, M. E., & DiVesta, F. J. (1986). The effects of conscious controlled verbalization of a cognitive strategy on transfer in problem solving. *Memory and Cognition, 14,* 281-285.

Austin, J. (1996). *Solving performance problems: Organizational troubleshooting in expert management consultants and experienced managers.* Unpublished doctoral dissertation, Florida State University, Tallahassee, FL.

Boreham, N. C. (1986). A model of efficiency in diagnostic problem solving: Implications for the education of diagnosticians. *Instructional Science, 15,* 191-211.

Bower, A. C., & King, W. L. (1967). The effect of number of irrelevant stimulus dimensions, verbalization, and sex on learning bi-directional classification rules. *Psychonomic Science, 8,* 453-454.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance, 11,* 1-27.

Bristow, D., & Fristoe, M. (1984). Learning of blissymbols and manual signs. *Journal of Speech and Hearing Disorders, 49,* 145-151.

Catania, A. C., Shimoff, E., & Matthews, B. A. (1989). An experimental analysis of rule-governed behavior. In S. C. Hayes (Ed.), *Rule-governed behavior: Cognition, contingencies, and control* (pp. 119-150). New York: Plenum.

Ceci, S. J., & Bruck, M. (1993). Suggestibility of the child witness: A historical review and synthesis. *Psychological Bulletin, 113,* 403-439.

Chase, W. G., & Ericsson, K. A. (1981). Skilled memory. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 141-189). Hillsdale, NJ: Erlbaum.

Chase, W. G., & Ericsson, K. A. (1982). Skill and working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 16, pp. 1-58). New York: Academic Press.

Critchfield, T. S., & Perone, M. (1990). Verbal self-reports of delayed matching to sample by humans. *Journal of the Experimental Analysis of Behavior, 53,* 321-344.

Crutcher, R. J., Ericsson, K. A., & Wichura, C. A. (1994). Improving the encoding of verbal reports by using MPAS: A computer-aided encoding system. *Behavior Research Methods, Instruments, and Computers, 26,* 167-171.

de Groot, A. D. (1965). *Thought and choice in chess.* The Hague: Mouton.

Delaney, P. F. (1997, May). Active behaviors in "storage" of information. In J. Austin (Chair), *Human memory and remembering behavior.* Symposium conducted at the 23rd annual convention of the Association for Behavior Analysis, Chicago.

Delaney, P. F., & Austin, J. (1998). Memory as behavior: The importance of acquisition and remembering strategies. *The Analysis of Verbal Behavior, 15,* 75-91.

Devany, J. M., Hayes, S. C., & Nelson, R. O. (1986). Equivalence class formation in language-able and language-disabled children. *Journal of the Experimental Analysis of Behavior, 46,* 243-257.

Duncker, K. (1926). A qualitative (experimental and theoretical) study of productive thinking (solving of comprehensible problems). *Pedagogical Seminary, 33,* 642-708.

Ericsson, K. A., & Delaney, P. F. (in press). Long-term working memory as an alternative to capacity models of working memory in everyday skilled performance. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control.* New York: Cambridge University Press.

Ericsson, K. A., & Polson, P. G. (1988). Memory for restaurant orders. In M. T. H. Chi, R. Glaser, & M. Farr

(Eds.), *The nature of expertise* (pp. 23-70). Hillsdale, NJ: Erlbaum.

Ericsson, K. A., & Simon, H. A. (1987). Verbal reports on thinking. In C. Faerch & G. Kasper (Eds.), *Introspection in second language research* (pp. 24-53). Philadelphia: Multilingual Matters, Ltd.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (rev. ed.). Cambridge, MA: MIT Press.

Gagne, R. H., & Smith, E. C. (1962). A study of the effects of verbalization on problem solving. *Journal of Experimental Psychology, 63*, 12-18.

Hayes, L. J. (1994). Thinking. In S. C. Hayes, L. J. Hayes, M. Sato, & K. Ono (Eds.), *Behavior analysis of language and cognition* (pp. 149-164). Reno, NV: Context Press.

Hayes, S. C. (1986). The case of the silent dog – verbal reports and the analysis of rules: A review of Ericsson and Simon's *Protocol Analysis: Verbal Reports as Data*. *Journal of the Experimental Analysis of Behavior, 45*, 351-363.

Hayes, S. C., White, D., & Bissett, R. T. (1998). Protocol analysis and the "silent dog" method of analyzing the impact of self-generated rules. *The Analysis of Verbal Behavior, 15*, 57-63.

Hendrix, G. (1947). A new clue to transfer of training. *Elementary School Journal, 48*, 197-208.

Hodges, P., & Schwethelm, B. (1984). A comparison of the effectiveness of graphic symbol and manual sign training with profoundly retarded children. *Applied Psycholinguistics, 5*, 223-253.

Karpf, D. (1972). Thinking aloud in human discrimination learning (Doctoral dissertation, State University of New York at Stony Brook, 1972). *Dissertation Abstracts International, 33*, 6111-B. (University Microfilms No. 73-13625)

Karpf, D., & Levine, M. (1971). Blank-trial probes and introtacts in human discrimination learning. *Journal of Experimental Psychology, 90*, 51-55.

Lane, S. D., & Critchfield, T. S. (1996). Verbal reports of emergent relations in a stimulus equivalence procedure. *Journal of the Experimental Analysis of Behavior, 65*, 355-374.

Lindsley, O. R. (1996). Is fluency free-operant response-response chaining? *The Behavior Analyst, 19*, 211-224.

Loftus, E. F. (1996). Memory distortion and false memory creation. *Bulletin of the American Academy of Psychiatry and the Law, 24*, 281-295.

Loftus, E. F., & Hoffman, H.G. (1989). Misinformation and memory: The creation of new memories. *Journal of Experimental Psychology: General, 118*, 100-104.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*, 585-589.

Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals, 25*, 720-725.

Lowenkron, B. (1991). Joint control and the generalization of selection-based verbal behavior. *The Analysis of Verbal Behavior, 9*, 121-126.

Manning, C. G., & Loftus, E. F. (1996). Eyewitness testimony and memory distortion. *Japanese Psychological Research, 38*, 5-13.

Michael, J. (1985). Two kinds of verbal behavior plus a possible third. *The Analysis of Verbal Behavior, 3*, 1-4.

Michael, J. (1993). *Concepts and principles of behavior analysis*. Kalamazoo, MI: Society for the Advancement of Behavior Analysis.

Nelson, R. O. (1977). Assessment and therapeutic functions of self-monitoring. In M. Hersen, R. M. Eisler, & P. Miller (Eds.), *Progress in behavior modifcation* (Vol. 5, pp. 263-308). New York: Academic Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259.

Perone, M. (1988). Laboratory lore and research practices in the experimental analysis of human behavior: Use and abuse of subjects' verbal reports. *The Behavior Analyst, 11*, 71-75.

Potter, B. (Chair). (1997, May). *The use of protocol analyses in research*. Panel discussion conducted at the 23rd annual convention of the Association for Behavior Analysis, Chicago.

Potter, B., Huber, S., & Michael, J. (1997). The role of mediating verbal behavior in selection-based responding. *The Analysis of Verbal Behavior, 14*, 41-56.

Reber, A. S., & Lewis, S. (1979). Implicit learning: An analysis of the form and structure of a body of tacit knowledge. *Cognition, 5*, 333-361.

Rommetveit, R. (1960). Stages in concept formation and levels of cognitive functioning. *Scandinavian Journal of Psychology, 1*, 115-124.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition 17*, 759-769.

Schweiger, D. M. (1983). Is the simultaneous verbal protocol a viable method for studying managerial problem solving and decision making? *Academy of Management Journal, 26*, 185-192.

Shimoff, E. (1984). Post-session questionnaires. *Experimental Analysis of Human Behavior Bulletin, 2*, 1.

Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech and Hearing Research, 14*, 5-13.

Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior, 37*, 5-22.

Simon, H. A. (1979). *Models of thought* (Vol. 1). New Haven, CT: Yale University Press.

Skinner, B. F. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.

Skinner, B. F. (1974). *About behaviorism*. New York: Knopf.

Sorabji, R. (1972). *Aristotle on memory*. Providence, RI: Brown University Press.

Storms, M. D. (1973). Videotape and the attribution process: Reversing actors' and observers' points of view. *Journal of Personality and Social Psychology, 27*, 165-175.

Sundberg, C. T., & Sundberg, M. L. (1990). Comparing topography-based verbal behavior with stimulus selection-based verbal behavior. *The Analysis of Verbal Behavior, 8*, 31-41.

Titchener, E. B. (1912). The schema of introspection. *American Journal of Psychology, 23*, 485-508.

Watson, J. B. (1920). Is thinking merely the action of language mechanisms? *British Journal of Psychology, 11*, 87-104.

Wilder, L., & Harvey, D. J. (1971). Overt and covert verbalization in problem solving. *Speech Monographs, 38*, 171-176.

Wulfert, E., Dougher, M. J., & Greenway, D. E. (1991). Protocol analysis of the correspondence of verbal behavior and equivalence class formation. *Journal of the Experimental Analysis of Behavior, 56*, 489-504.

APPENDIX

Standard Think-Aloud Instructions

In this experiment, we are interested in what you think about when you find answers to some questions that I am going to ask you to answer. In order to do this I am going to ask you to THINK ALOUD as you work on the problem given. What I mean by think aloud is that I want you to tell me EVERYTHING you are thinking from the time you first see the question until you give an answer. I would like you to talk aloud CONSTANTLY from the time I present each problem until you have given your final answer to the question. I don't want you to try to plan out what you say or try to explain to me what you are saying. Just act as if you are alone in the room speaking to yourself. It is most important that you keep talking. If you are silent for any long period of time I will ask you to talk. Do you understand what I want you to do? (from Ericsson & Simon, 1993, p. 378)